Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 4 (2011) 1119–1128

Procedia Computer Science

International Conference on Computational Science, ICCS 2011

The ACGT project in retrospect: Lessons learned and future outlook

Anca Bucur[a], Stefan Ruping[b], Thierry Sengstag[c], Stelios Sfakianakis[d], Manolis Tsiknakis[d,*], Dennis Wegener[b]

[a] *Philips Research Europe, The Netherlands*
[b] *Fraunhofer IAIS, Germany*
[c] *RIKEN Yokohama Institute, Japan*
[d] *Biomedical Informatics Laboratory, ICS-FORTH, Greece*

Abstract

The objective of the ACGT (Advancing Clinico-Genomic Trials on Cancer: Open Grid Services for improving Medical Knowledge Discovery, www.eu-acgt.org) project that has recently concluded successfully was the development of a semantically rich infrastructure facilitating seamless and secure access and analysis, of multi-level clinical and genomic data enriched with high-performing knowledge discovery operations and services in support of multi-centric, post-genomic clinical trials. In this paper we describe the way the ACGT consortium has approached important challenges in the design and the execution of the clinical trials such as the issues of data integration, semantics based data fusion, data processing and knowledge extraction, privacy and security, etc. Furthermore we provide a number of key "lessons learned" during the process and give directions for further developments in the future.

Keywords: bioinformatics, data integration, Grid, clinical trials

1. Introduction

Recent advances in methods and technologies in molecular biology have resulted in an explosion of information and knowledge about cancer and its treatment. As a result, our ability to characterize and understand the various forms of cancer is growing exponentially. Information arising from post-genomics research and combined genetic and clinical trials on one hand, and advances from high-performance computing and informatics on the other, is rapidly providing the medical and scientific community with an enormous opportunity to improve prognosis of patients with cancer by individualizing treatment. To achieve this goal, a unifying platform is needed that has the capacity to process this huge amount of multi-level and heterogeneous data in a standardized way. Multi-level data collection within clinico-genomic trials and interdisciplinary analysis by clinicians, molecular biologists and others involved in life science is mandatory to further improve the outcome of cancer patients. It is essential to merge the research results of biomolecular findings, imaging studies and clinical data of patients and to enable users to easily join, analyze and share even great amounts of data.

*Corresponding author
Email address: tsiknaki@ics.forth.gr (Manolis Tsiknakis)

Selection and/or peer-review under responsibility of Prof. Mitsuhisa Sato and Prof. Satoshi Matsuoka
doi:10.1016/j.procs.2011.04.119

An important challenge in carrying out post-genomic bio-medical research is therefore to efficiently manage and retrieve all relevant data from many heterogeneous sources. A post-genomic clinical trial involves the collection, storage and management of a wide variety of data, including: clinical data collected on Case Report Forms (e.g. symptoms, histology, administered treatment, treatment response), imaging data, genomic data, pathology data and other lab data. Next to that, access to many external sources of data and knowledge is required. These store information about gene and protein sequences, pathways, genomic variation, microarray experiments, medical literature, etc. Seamless access to all these data repositories would greatly facilitate research.

Furthermore the state-of the-art clinical research requires an array of data manipulation, visualization, statistical, and knowledge extraction tools in order to gain insight into the meaning of the data and answer the specific research questions posed. To provide a functional and user-friendly suite of tools in a coherent platform it is of utmost importance that the development of such a platform is user-driven and evaluated by end users right from the planning and development phase. Such tools and software should also be based on the user's needs and have to be in accordance with ethical and legal requirements of the European Community.

The ACGT (Advancing Clinico-Genomic Trials on cancer: Open Grid Services for improving Medical Knowledge Discovery)[1] was an Integrated Project (IP) funded in the 6th Framework Program of the European Commission that aimed at providing solutions to the challenges described above. The ACGT technological platform is based on an ontology-driven, semantic grid services infrastructure that enables the efficient execution of discovery-driven analytical workflows in the context of multi-centric, post-genomic clinical trials. The ultimate objective of the ACGT project was the development of a secure semantic grid services infrastructure which will (a) facilitate seamless and secure access to heterogeneous, distributed multilevel databases; (b) provide a range of semantically rich re-usable, open tools for the analysis of such integrated, multilevel clinico-genomic data; (c) achieve these results in the context of discovery-driven (eScience) workflows and dynamic VOs; and (d) fulfill these objectives while complying with existing ethical and legal regulations. In this paper we describe the way the ACGT consortium approaches important challenges in the design and the execution of the clinical trials such as the issues of data integration, semantics based data fusion, data processing and knowledge extraction, privacy and security, etc. Furthermore we provide a number of key "lessons learned" during the process and give directions for further developments in the future.

1.1. Architecture and key components

Because of the complexity and the diversity of the user requirements and the biomedical domain in general a multidisciplinary and multi-paradigm approach was followed, according to the following technologies and standards: Service Oriented Architecture (Web Services [2]), the Grid [3], and the Semantic Web [4]. These underlying technologies work complementary to each other, providing their benefits in different aspects of the technological infrastructure. The Grid provides the computational and data storage infrastructure, the general security framework, the virtual organization abstraction and relevant user management mechanisms etc. The machine to machine communication is performed via XML programmatic interfaces over web transport protocols, which are commonly referred as Web Services interfaces. Finally the Semantic Web adds the universal data modeling technology in terms of the RDF abstract model, the knowledge representation mechanisms through the means of OWL ontologies, the implementation-neutral query facilities with the SPARQL query language and the associated query interfaces, etc.

The adopted architecture for ACGT is shown in Fig. 1. A layered approach has been followed for providing different levels of abstraction and a classification of functionality into groups of homologous software entities. In this approach we consider the security services and components to be pervasive throughout the ACGT platform, so as to cater both for the user management, access rights management and enforcement, and trust bindings that are facilitated by the grid and domain specific security requirements like pseudonymization.

The sensitivity of the patient data requires a strong security framework to provide enough safety nets in order to maintain privacy, confidentiality, and integrity. The grid middleware already supports much of the necessary infrastructure, in terms of certificate based Grid Security Infrastructure (GSI)[5], the Virtual Organization (VO) abstraction and the user credential management, and the Grid Authorization Services (GAS). In ACGT this "system level" security is complemented by "domain specific" mechanisms like pseudomymization that permits the identification of patient specific information without revealing the true person identity [6]. All data is anomymized before their entry in the ACGT domain and, even during their analysis, all the processing tasks are audited and authorized based on the end users' identity and access rights as specified in the context of a clinical trial.

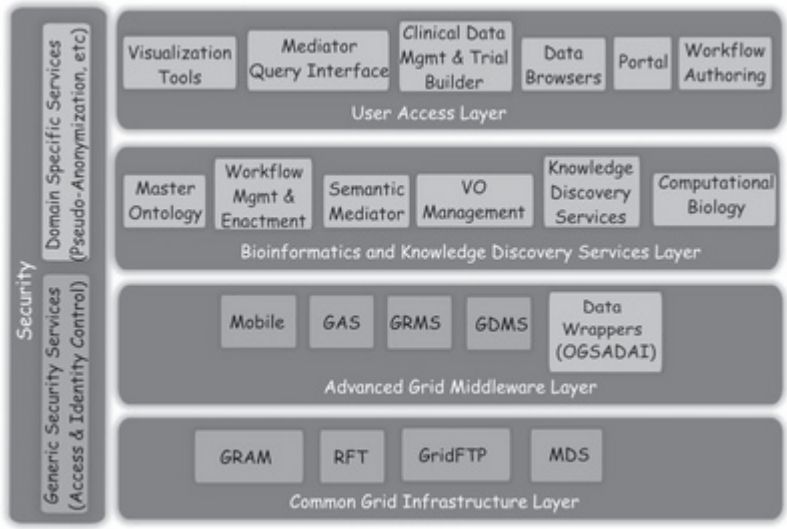

Fig. 1: The ACGT Architecture

Apart from the security requirements, the Grid infrastructure and other services are located in the first (lowest) two layers: the Common Grid Layer and the Advanced Grid Middleware Layer. The upper layer is where the user access services, such as the ACGT Portal and the visualization tools, reside. Finally, the Bioinformatics and Knowledge Discovery Services are the workhorse of ACGT and the corresponding layer is where the majority of the ACGT specific services lie. Some of the most important components and services of this layer are described below.

1.1.1. The ACGT Master Ontology on Cancer

The ACGT Master Ontology on Cancer (ACGT MO) has been developed with the goal of building a consistent semantic framework for describing the domain of post-genomic clinical trials on cancer [7]. This framework is the basis of the semantic interoperability to connect the different services and data sources in the ACGT Platform: it is the "lingua franca" for the integration, analysis, and synthesis of data. It is written in OWL-DL – which allows automatic reasoning e.g. for consistency checking – and contains more than 1600 classes and near 300 properties.

1.1.2. Clinical Trial Builder: ObTima

ObTiMA is an ontology-based system for creating and conducting clinical trials on cancer [8]. The system includes a graphical Trial Builder and facilitates the trial chairman in the design of the Case Report Forms (CRFs) to be used for each treatment step. The design of the CRFs is based on the ACGT MO which means that ontology compliance is "built in" and interoperability or syntactic transformation is to great extent guaranteed. The data collected in the trial is stored in trial databases in the Patient Data Management System, which is the other component of ObTiMA that is setup automatically in such a way that a medical clinician can collect the patient data during the trial (Fig. 2).

1.1.3. Data Access Services

The Data Access Layer of the ACGT platform consists of the Database Wrappers and the Semantic Mediator [9]. The database wrappers deal with the syntactic heterogeneities, offering a uniform query interface (SPARQL) to data resources. On the other hand the Semantic Mediator tackles the semantic heterogeneities – i.e. offering a common data model for accessing the data resources exposed by the wrappers and performing query translation from the global schema to the local schemata of the integrated databases. The ACGT MO has been adopted as the global model exposed to the clients of the Data Access Layer through a single SPARQL endpoint. The Semantic Mediator translates the query expressed in terms of the Master Ontology to the language of the Data Access Service. Fig. 3 shows an example of a simple query called "Patient Weights" before and after the mediation process.

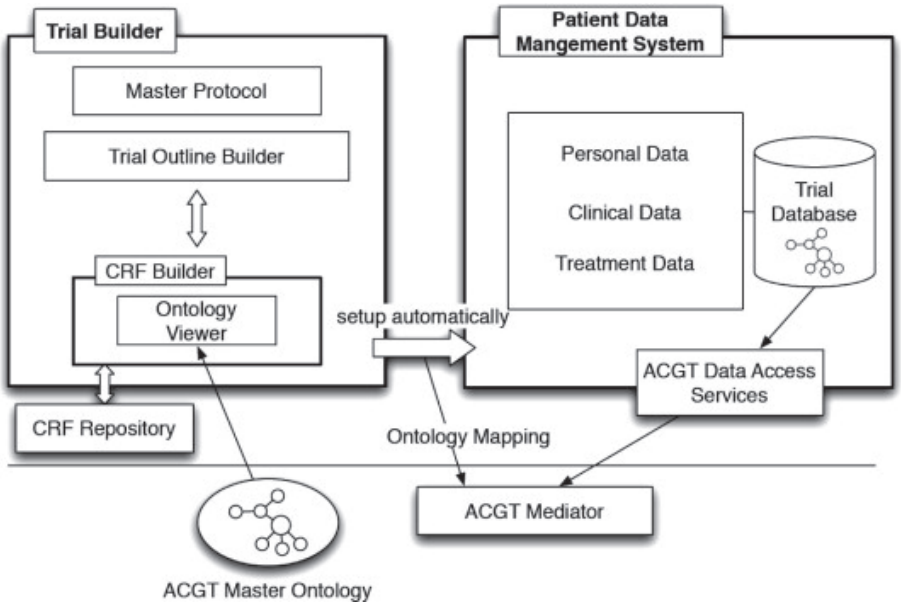

Fig. 2: The components of ObTiMA

1.1.4. KDD Tools

The ACGT Platform comprises a series of knowledge discovery tools for analyzing and extracting useful information from data collected in a clinical trial. With an abundance of such tools available freely it was decided that the core question does not relate with the development of new tools, but rather with how to seamlessly integrate existing toolkits.

The R language has been adopted as the prime tool for carrying out statistical analysis of the data. The GridR tool [10] allows to seamlessly execute R jobs and addresses the needs of the users by enabling them to use their standard analysis tools in the context of large, distributed eScience systems. In ACGT GridR plays a dual role: on one hand it can be used interactively, giving the users access to the whole ACGT environment, on the other hand it is deployed as a data-analysis service exposing a Web Service interface for the execution of scripts incorporated in scientific workflows. This facilitates the efficient development, execution, and re-use of analytical solution without the need for any knowledge about the underlying architecture on the side of the analyst. Additionally integration

```
PREFIX acgt:     <http://www.ifomis.org/acgt/1.0#>
PREFIX xsd:      <http://www.w3.org/2001/XMLSchema#>
PREFIX bfo_span: <http://www.ifomis.org/bfo/1.1/span#>
PREFIX bfo_snap: <http://www.ifomis.org/bfo/1.1/snap#>
PREFIX ro:       <http://www.ifomis.org/obo/ro/1.0#>

SELECT ?weightValue
WHERE {

  ?patient     acgt:hasWeight     ?weight .
  ?weight      acgt:hasFloatValue ?weightValue .
  ?patient     a                  acgt:HumanBeing .
  ?weight      a                  acgt:Weight .
  ?weightValue a                  xsd:float .

}
```

```
PREFIX vocab: &lt;http://obtima.ibmt.fhg.de/1.0#&gt;
PREFIX xsd: &lt;http://www.w3.org/2001/XMLSchema#&gt;

SELECT ?v
WHERE {
  ?patient vocab:Patient_CRF      ?crf .
  ?crf     vocab:CRF_Item         ?item .
  ?item    vocab:item_entry_value ?v .
  ?patient a                      vocab:patient .
  ?crf     a                      vocab:crf_instance .
  ?item    a                      vocab:item_entry .

  ?crf vocab:ci_crf_template_fk ?rest0 .
  FILTER ( ?rest0 = "57"^^xsd:integer )

  ?item vocab:ie_item_template_fk ?rest1 .
  FILTER ( ?rest1 = "327"^^xsd:integer )
}
```

Fig. 3: (a) A “Patient Weights” query for the Semantic Mediator; (b) The query after the translation as submitted to a trial database

with the Biomoby open source registry of tools [11] allows the use of even a larger set of domain specific analyses to performed in the context of the ACGT supported clinical trials.

1.1.5. The ACGT Workflow environment

To assist bioinformaticians in building their complex scientific workflows, a Workflow Editor and Enactment Environment has been implemented [12]. This environment is accessible through the ACGT Portal and includes a suite of tools that allow users to combine different web services into complex workflows. An intuitive user interface permits searching registered services – e.g. GridR scripts – and retrieving data through the Data Access Layer. These elements can then be combined and orchestrated to produce the expected workflow. The designed workflows can be stored in a user's specific area and later retrieved and edited. Workflows are executed on a remote machine or even in a cluster of machines in the Grid so there is no burden imposed on the user's local machine. The publication, annotation, and sharing of the workflows are also supported so that the user community can exchange information benefitting from each other's research.

2. Evaluation in an exemplary scenario

Evaluation of the ACGT platform has been done in the context of clinically oriented data analysis scenarios. One such scenario is the so called "Multi Center Multi Platform" (MCMP) scenario, that aimed at demonstrating the utility of the platform as an information system to exploit data in the context of a clinical trial. In this scenario biopsies are collected from patients registered in two centers and each center is using a different microarray platform, namely Affymetrix and Illumina, to measure genes expression in the samples. In addition, the classical clinical parameters associated to each patient are available in a trial database. This specifies a multi-centric and multi-platform study, with only one microarray per patient. All patient private data were anonymized prior to their integration in the ACGT environment. The whole scenario has been realized by the means of a scientific workflow shown in Fig. 4. The process begins by retrieving microarray experiments that are stored as files in the Grid file system and preprocessing them in two parallel branches based on the platform used. A feature selection stage is then performed in each of these branches to extract the most informative genes. At the final step, the results of the two parallel subprocesses are combined in an analysis/biomarker discovery task that also uses the results returned by the Semantic Mediator based on the MO-expressed query for the patients' clinical data. This workflow, which implements a methodology linking microarrays and classical clinical data and used in biomarker discovery, illustrates the ability of the ACGT platform to repeat complex analyses on an evolving population of patients, ranging from data retrieval and integration, normalization, to analysis and results presentation.

3. Lessons learnt, experiences gained

The exploitation of a broad and complex European Commission collaborative project is a multi-dimensional problem with many challenges as well as opportunities. The experience has provided lessons that could provide additional guidance to others who might decide to work in the broad areas covered by ACGT. In this section we present these lessons learned covering the development and exploitation of specific modules) as well as broader aspects of the project as a whole.

3.1. Technological Approach

The ACGT infrastructure is a complex technical infrastructure to which many different partners have contributed. Such large technical collaborations can only succeed when sufficient attention is given to using as much as possible industry standards. ACGT has done this; however, the infrastructure produced is still rather monolithic due to the pervasive use of its Grid middleware. The big advantage of relying on the existing middleware is that it provides standard solutions to common complex tasks (e.g. delegation of rights, orchestration, resource allocation, etc). In hindsight, a technological undertaking of the scale of ACGT (with so many contributors) might benefit from a more lightweight solution. This means that one is burdened with a number of tasks otherwise covered by the middleware, but experience has shown that not all functionality offered by the complex middleware is strictly necessary in the trial application domain. On the other hand a lightweight architecture can evolve more dynamically (e.g. adopting new technology)

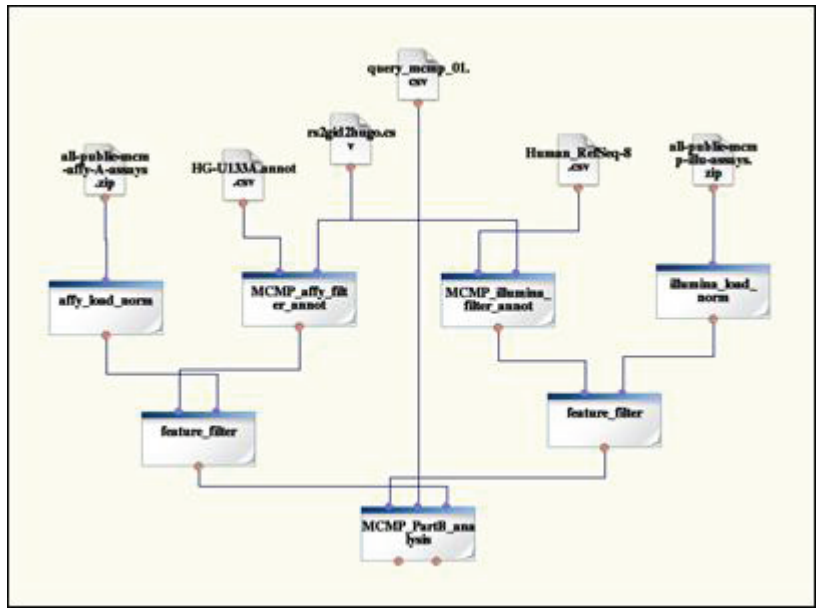

Fig. 4: The MCMP workflow

over time. Cooperation with a large number of people can also be more efficient if the focus of collaboration is on interfacing, interoperability and integration in a more loosely coupled architecture. Quality control and acceptance testing in view of compliance could also benefit from this approach. It should also be noted, that although a more lightweight (loosely coupled) architectural approach can in practice easily evolve into "a new" heavyweight middleware (as more functionality is shifted to centrally maintained and deployed proprietary components), the potential reward makes the approach worth investigating.

3.2. Management of clinical research information and building comprehensive datasets

One of the main challenges in carrying out post-genomic research is to efficiently have access to all relevant data. In the context of clinical trials, this data may be scattered geographically and resides in a variety of different systems. The data typically comprises clinical data collected on Case Report Forms (e.g. symptoms, histology, administered treatment, treatment response), imaging data (e.g. X-Ray, CT, MR, Ultrasound), genomic data (e.g. microarray data), pathology data and other lab data. Next to that there are many public biomedical databases that are relevant. These store information about gene and protein sequences, pathways, genomic variation, microarray experiments, medical literature, tumor antigens, protein domains, metabolites, etc.

In ACGT, users can query a data resource and obtain its RDF schema when it is available within the ACGT platform. If the data resource is fully integrated into the ACGT platform, the queries are expressed in terms of the ACGT Master Ontology, otherwise the queries are expressed in terms of the data resources local ontology.

A relevant challenge is dealing with the generality of the ACGT MO versus the specificity of a (legacy) database that we need to integrate. The database schema is developed with a specific concrete goal in mind, while the ACGT MO intends to be very generic in an attempt to represent the domain of cancer research and management. When trying to explore a data resource by formulating SPARQL queries using the ACGT MO as a guide, because the ontology captures such a wide domain, it is very easy to specify a constraint on a term that is not used in the actual data resource, resulting in an empty solution as response.

Another challenge is related to enabling users to properly understand and use a new data resource. When first encountering a (potentially) useful data resource, users often try to get to understand the data resource by exploring it. This can be quite cumbersome when the layout of the underlying data source is according to a generic database scheme. The ObTima database schema is a generic clinical trial schema, almost similar to a meta-schema. For the uninitiated user, the generic layout is impossible to understand on its own. To understand the setup of the specific

clinical trial in the generic layout, one has to descend to the instance level of the data. The cause of this problem is the lack of a (preferably standardized) description of the content of the data resource aimed at the (human) user. In addition, there is currently no tool available to help to browse conveniently through the data contained in a data resource.

The idea of integrating clinical practice and clinical research gets more and more traction. This will have large consequences such as a higher innovation pace in patient care, and improved patient safety (e.g. due to less duplication of patient data entry). The context of the ACGT project is clinical research, which has an effect on the way data is recorded. In clinical practice, the primary and often only purpose for data recording is direct patient care (on a per-patient basis), and a lot of data is recorded in free text format. Due to the clinical trial requirements (to allow for a correct and efficient analysis), the data collected is typically structured. When generalizing the ACGT effort to include data collected in clinical practice, it should be investigated how to integrate free text resources into the environment (this includes natural language processing, and semantic integration of the results).

Part of the evaluation of the ACGT results, we have investigated how the expertise, and potentially also the tools, developed in ACGT could be used to support a large real-life multi-centric clinical trials program, such as NeoBIG, the new research program of the Breast International Group (`http://www.breastinternationalgroup.org/`). To suit the ACGT scope, our focus was on the IT needs of the NeoBIG research program, specifically with respect to secure privacy-preserving data management and sharing as these are issues at the core of ACGT.

During the study, we have concluded that there is a lot of ACGT expertise that could be used for the NeoBIG data sharing platform, especially with respect to data storage, management and sharing, and with respect to privacy and security. At the same time, we understood that while accessing external data out of heterogeneous repositories is highly relevant, there is also very high value in supporting the research community to define common methodologies concerning data management and to build and share comprehensive datasets including all the wealth of data collected in the trials. The advantage of building such consolidated data sets under a single authority in charge of their maintenance is that coherence, adherence to common methodologies, standards and ontologies, and availability can be ensured. While a solution maintaining all the data at the institutions generating that data is feasible and provides flexibility and scalability, it does not guarantee adherence to the same methodologies, common data models and standards, or long term maintenance. This makes enabling (long term) use of the combined datasets more difficult.

3.3. Semantic Data Integration

Developing a semantic integration layer for biomedical databases resulted in a quite complex task. The biomedical domain, and more specifically, the cancer-related clinical trials domain, evolves at a surprisingly high rate. We found out that, during the four and a half years that the project lasted, new requirements appeared, or initial requirements had to suffer modifications, simply because the biomedical field had new needs. We found it was crucial to adopt highly flexible designs for our tools, so that they could be adapted to the new needs without requiring deep changes in the code. Due to the extension and the rich feature nature of the tools to develop, it was necessary to adopt third party tools or APIs available for the research community. Special care must be taken when selecting what tools to use, since lack of documentation or support can seriously affect development.

3.4. Data Analysis Processes

The ACGT system strives to integrate all steps from the collection and management of various kinds of data in a trial up to the statistical analysis by the researcher. However, the more powerful these environments become, the more important it is to guide the user in the complex task of constructing appropriate workflows. This is particularly true for the case of workflows which encode a data mining tasks, which are typically much more complex and in a more constant state of frequent change than workflows in business applications. From the ACGT project, we learned that [13]:

1. The construction of data mining workflows is an inherently complex problem when it is based on input data with complex semantics, as it is the case in clinical and genomic data.
2. Because of the complex data dependencies, copy and paste is not an appropriate technique for workflow reuse.
3. Standardization and reuse of approaches and algorithms works very well on the level of services, but not on the level of workflows. While it is relatively easy to select the right parameterization of a service, making the right connections and changes to a workflow template is quickly getting quite complex, such that user finds it easier to construct a new workflow from scratch.

4. Workflow reuse only occurs when the initial creator of a workflow describes the internal logic of the workflow in detail. However, most workflow creators avoid this effort because they simply want to "solve the task at hand".

Thus, the situation of having a large repository of workflows to choose the appropriate one from, which is often assumed in existing approaches for workflow recommendation systems, may not be very realistic in practice.

While a multitude of tools for data mining, bioinformatics, and statistics on clinical data exists, the question of quality control and standardization, as well as ease of use and reusability, remains largely unanswered yet.

In the business process management community, pattern-based approaches for facilitating reuse have been proposed, e.g. by the definition of workflow patterns that describe the control-flow perspective of workflow systems [14] based on business process modeling formalisms (including BPMN) and business process execution languages (including BPEL). Next to BPMN, also UML 2.0 Activity Diagrams have been proposed as notations that are compatible with workflow patterns [15]. Process patterns have many advantages [16]: BPM processes serve as both the specification and the source code. The modeled processes become the solutions deployed and provide a simple communication tool between end-users, business analysts, developers and the management.

To bridge the gap between providing single tools and the knowledge of how to make good use of these tools to properly answer a biomedical research question, we propose support for a pattern based approach for data mining [17]. A data mining pattern can be viewed as template for a concrete data mining workflow. It acts as a formalized common language between data mining experts and biomedical researchers. It contains not only a reusable workflow template, but also a description of the requirements, assumptions, and steps that need to be fulfilled to apply the generic data mining solution to a specific problem. It thus describes a data mining solution not only syntactically (which steps need to be executed) but also semantically (when does it make sense to apply the solution and what result can one expect?). In future work, we intend to demonstrate that data mining process patterns provide a simple technique to shorten the learning curve and improve productivity and quality of the analysis processes, as they are simple to understand, learn and apply immediately.

3.5. Security and Data Protection

Compliance to trial related legislation, especially to the data protection laws, is a critical success factor for any research-network. ACGT has expended considerable effort in order to automate achieving this compliance, for example through the founding of Center for Data Protection (`https://cdp.custodix.com/`) for establishing necessary contracts and with tools such as CAT (Custodix Anonymisation Tool) that ease the anonymization of the data.

ACGT has shown us, that investing in achieving compliance by default (without specific effort or much expertise from the end-users) with fixed procedures is a must for long term success and smooth operations. In view of the experience, one could suggest that this requirement should be valued very high when making technological choices.

3.6. User empowerment and Community building

In terms of deploying the ACGT infrastructure and resources in end-user environments consortium exploitation efforts have highlighted a number of issues that merit specific attention by similar efforts in the future. They are the following:

- Not enough infrastructure and services in place: the lack of enough related end-user oriented modules in ACGT is believed to be one of the factors that have hindered its deployment in actual working environments. In particular, the 'completeness' and consistency of the available set of end user services has not been what is required to support existing workflows and tasks in end user partner environments. While considerable effort was spent in the last year to define scenarios that represent actual tasks and workflows that make sense to end users (doctors, researchers, etc.). these were still 'fragmented' being able to support only simple operations.

- Opt-in versus Opt-out and the anti spam laws: The ACGT Management Board decided to adopt a conservative approach in all its mailing and communication efforts. Opt-in rather than opt-out policies were followed. As a consequence, circulation and site visit figures were rather low. It is felt that the failure of the ACGT competition to attract sufficient interest was also partly due to this policy which should not be adopted by similar projects.

- Continued support of ACGT infrastructure: A number of contacted third parties expressed this concern at various points in time. While we are not able to ascertain to what degree this has acted as a true deterrent to adoption, this is certainly an important concern that must be addressed before successful uptake can be expected.
- Intellectual property: IP-issues tend to be a hindering factor in data exchange: On the one hand participating clinicians have serious reservations against sharing (raw) patient data as their possession is an important (and not always legally protected) factor in scientific competition. On the other hand patients' (sometimes economic) interests in the outcome of the research are not always sufficiently covered by trial setups and results' exploitation. ACGT has developed guidelines for bringing decision makers into the position to allow patients and clinicians proper participation in the exploitation process. This work can serve as a basis for European project managers in the E-health area to identify intellectual property issues in an early stage of the project's lifecycle

4. Conclusions

When launched back in 2006, the ACGT project aimed at providing clinical researchers with an infrastructure that would support the requirements of modern clinical trials. At the same time there were a number of projects that aim at developing grid-based infrastructure for post-genomic cancer clinical trials, the most advanced of which are NCI's caBIG (Cancer Biomedical Informatics Grid, `https://cabig.nci.nih.gov/`) in the USA and CancerGrid (`http://www.cancergrid.org/`) in the UK. The overall approach in those projects was somewhat different from the one in ACGT. In caBIG, the bottom-up, technology-oriented, approach was chosen, in which the focus was put on the integration of a large number of analysis tools but with weak concern on data privacy issues. CancerGrid on the other hand addresses the very needs of the British clinical community. In contrast, the ACGT project was focused on the development of a pan-european system that is driven by current demands from clinical practice. With two on-going international clinical trials actually conducted in the framework of the project, the approach was top-down, with clinicians' and biomedical data analysts' needs at the heart of all technical decisions, considering data privacy issues as central as data analysis needs.

In this user driven endeavor the technical concerns raised by the multiplicity and heterogeneity of user requirements demanded state of the art methodologies and technologies. In the ACGT work plan the employment of ontologies and metadata annotations and the realization of intelligent higher level services were the primary implementation targets. Five years after the launch of the project the consensus among the consortium is that we have made significant progress towards the initial objectives. Some of (mostly technological) decisions made may look suboptimal or have been superseded by today's offerings (e.g. the rise of cloud computing replacing the Grid) but the methodology and the core design principles remain valid. The security framework, the ontology driven approach, the emphasis on simplifying researchers' daily work focusing on user-friendliness, etc. are of paramount importance in similar endeavors in contemporary applied research and science.

On the other hand we acknowledge the fact that the ACGT platform is still lacking when considered for everyday use in a production level environment. Due to the very strict requirements for a production-level system, with available documentation and user support, commercial deployment and long term maintenance, we have concluded that current ACGT solutions cannot be directly used e.g. for NeoBIG, however some of them could be part of a further targeted solution. For these reasons we are continuing our efforts in the context of follow up R&D projects, such as the INTEGRATE, a project that aims to build an environment that supports a large and multidisciplinary biomedical community ranging from basic, translational and clinical researchers to the pharmaceutical industry to collaborate, share data and knowledge, and build and share predictive models for response to therapies, with the end goal of improving patient outcome. The infrastructure and tools developed by the INTEGRATE project will support BIG to promote in the clinical community new methodologies and define standards concerning the collection, processing, annotation and sharing of data in clinical research and improve the reproducibility of results of clinical trials.

Acknowledgements

The authors gratefully acknowledge the financial support of the European Commission for the Project ACGT, FP6/2004/IST-026996.

References

1. M. Tsiknakis, M. Brochhausen, J. Nabrzyski, J. Pucacki, S. Sfakianakis, G. Potamias, C. Desmedt, D. Kafetzopoulos, A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of postgenomic clinical trials on cancer, Information Technology in Biomedicine, IEEE Transactions on 12 (2) (2008) 205–217.
2. F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, S. Weerawarana, Unraveling the Web services web: an introduction to SOAP, WSDL, and UDDI, Internet Computing, IEEE 6 (2) (2002) 86–93.
3. I. Foster, The grid: computing without bounds., Scientific American 288 (4) (2003) 78–85.
4. N. Shadbolt, W. Hall, T. Berners-Lee, The semantic web revisited, Intelligent Systems, IEEE 21 (3) (2006) 96–101.
5. V. Welch, F. Siebenlist, I. Foster, J. Bresnahan, K. Czajkowski, J. Gawor, C. Kesselman, S. Meder, L. Pearlman, S. Tuecke, Security for grid services, in: High Performance Distributed Computing, 2003. Proceedings. 12th IEEE International Symposium on, IEEE, 2003, pp. 48–57.
6. B. Claerhout, N. Forgó, T. Krügel, M. Arning, G. De Moor, A Data Protection Framework for Transeuropean genetic research projects, Studies in health technology and informatics 141 (2008) 67.
7. M. Brochhausen, A. Spear, C. Cocos, G. Weiler, L. Martín, A. Anguita, H. Stenzhorn, E. Daskalaki, F. Schera, U. Schwarz, et al., The ACGT Master Ontology and its Applications-Towards an Ontology-Driven Cancer Research and Management System, Journal of Biomedical Informatics 44 (1) (2011) 8 – 25.
8. G. Weiler, M. Brochhausen, N. Graf, F. Schera, A. Hoppe, S. Kiefer, Ontology based data management systems for post-genomic clinical trials within a european grid infrastructure for cancer research, in: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, 2007, pp. 6434 –6437.
9. L. Martín, A. Anguita, G. de la Calle, M. García-Remesal, J. Crespo, M. Tsiknakis, V. Maojo, Semantic data integration in the European ACGT project., in: AMIA Annu.Symp.Proc, 2007, p. 1042.
10. D. Wegener, T. Sengstag, S. Sfakianakis, S. Rüping, A. Assi, GridR: An R-based tool for scientific data analysis in grid environments, Future Gener. Comput. Syst. 25 (2009) 481–488.
11. M. Wilkinson, M. Links, BioMOBY: an open source biological web services proposal, Briefings in bioinformatics 3 (4) (2002) 331.
12. S. Sfakianakis, L. Koumakis, G. Zacharioudakis, M. Tsiknakis, Web-based authoring and secure enactment of bioinformatics workflows, Grid and Pervasive Computing Conference, Workshops at the 0 (2009) 88–95.
13. D. Wegener, S. Rüping, Re-using Data Mining in Business Processes – A Pattern–based Approach, in: Proceedings of the ECML/PKDD Workshop on Third-Generation Data Mining: Towards Service-Oriented Knowledge Discovery (SoKD'10), 2010, pp. 25 – 30.
14. N. Russell, W. van der Aalst, N. Mulyar, Workflow Control-Flow Patterns: A Revised View, BPM Center Report BPM-06-22.
15. N. Russell, W. M. P. van der Aalst, A. H. M. ter Hofstede, P. Wohed, On the suitability of uml 2.0 activity diagrams for business process modelling, in: Proceedings of the 3rd Asia-Pacific conference on Conceptual modelling - Volume 53, APCCM '06, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 2006, pp. 95–104.
16. D. Atwood, BPM Process Patterns: Repeatable Design for BPM Process Models, BP Trends May.
17. D. Wegener, S. Rüping, On Reusing Data Mining in Business Processes-A Pattern-based Approach, in: Business Process Management Workshops, Lecture Notes in Business Information Processing, Springer, 2010, proceedings of the 1st International Workshop on Reuse in Business Process Management (rBPM 2010).